Language Testing 2008 25 (2) 155–185

Rater types in writing performance assessments: A classification approach to rater variability

Thomas Eckes *TestDaF Institute, Germany*

Research on rater effects in language performance assessments has provided ample evidence for a considerable degree of variability among raters. Building on this research, I advance the hypothesis that experienced raters fall into types or classes that are clearly distinguishable from one another with respect to the importance they attach to scoring criteria. To examine the rater type hypothesis, I asked 64 raters actively involved in scoring examinee writing performance on a large-scale assessment instrument to indicate on a four-point scale how much importance they would attach to each of nine routinely used criteria. The criteria covered various performance aspects, such as fluency, completeness, and grammatical correctness. In a preliminary step, many-facet Rasch analysis revealed that raters differed significantly in their views on the importance of the various criteria. A two-mode clustering technique yielded a joint classification of raters and criteria, with six rater types emerging from the analysis. Each of these types was characterized by a distinct scoring profile, indicating that raters were far from dividing their attention evenly among the set of criteria. Moreover, rater background variables were shown to partially account for the scoring profile differences. The findings have implications for assessing the quality of large-scale rater-mediated language testing, rater monitoring, and rater training.

Keywords: classification, large-scale performance assessment, rater cognition, rater monitoring, rater variability

I Introduction

Numerous studies of rater behavior in performance assessment contexts have pointed to substantial degrees of unwanted rater variability – variability that is associated with characteristics of the raters and not with the performance of examinees (see, e.g., Engelhard, 1994; Bachman *et al*., 1995; Lumley & McNamara, 1995; Weigle, 1998;

Address for correspondence: Thomas Eckes, TestDaF Institute, University of Hagen, Feithstr. 188, 58084 Hagen, Germany; email: thomas.eckes@testdaf.de

 DOI:10.1177/0265532207086780

Congdon & McQueen, 2000; Engelhard & Myford, 2003; Eckes, 2005b; Schoonen, 2005). Moreover, rater training has been shown to be much less effective at reducing rater variability than expected; that is, raters typically remained far from functioning interchangeably even after extensive training sessions (Lumley & McNamara, 1995; Weigle, 1998; 1999; Hoyt & Kerns, 1999; Barrett, 2001) or after individualized feedback on their ratings (Elder *et al*., 2005).

Rater variability, or inconsistency between raters, can manifest itself in various forms (see Bachman & Palmer, 1996; McNamara, 1996; Weigle, 2002; Weir, 2005; Lumley, 2005). For example, raters may differ (a) in the degree to which they comply with the scoring rubric, (b) in the way they interpret criteria employed in operational scoring sessions, (c) in the degree of severity or leniency exhibited when scoring examinee performance, (d) in the understanding and use of rating scale categories, or (e) in the degree to which their ratings are consistent across examinees, scoring criteria, and performance tasks.

My focus in this research is on raters' interpretation of scoring criteria in a foreign language writing context. Specifically, I hypothesized that trained, experienced raters would differ in their interpretation of criteria used for essay scoring, and that these differences would contribute to rater variability. I further hypothesized that raters would fall into types (classes, clusters) that were clearly distinguishable from one another with respect to the importance raters attached to scoring criteria.

II Interpretation and use of scoring criteria

Scoring criteria play a crucial role in rater-mediated performance assessments. This is particularly evident in the case of multitrait or analytic scoring methods where assessments are made in relation to each of a number of criteria designed to represent central features of the language performance under consideration. Viewed from a rater cognition perspective, scoring criteria channel the ways in which raters perceive and evaluate concrete samples of language performance and, finally, come to assign scores to examinees (McNamara, 1996; Wolfe, 1997; Lumley, 2005).

Given the prominent status of scoring criteria, it is no wonder that much research in the field has focused on the range of performance features that raters actually attend to and use in the process of rating (see, e.g., Freedman, 1979; McNamara, 1990; Huot, 1993; Wolfe, 1997; DeRemer, 1998). Several studies have also examined differences

between particular groups of raters in the perception and use of criteria, especially differences between experienced or expert raters and inexperienced, untrained, or lay raters, thus taking up the issue of rater background (see, e.g., Brown, 1991; Pula & Huot, 1993; Hinkel, 1994; Schoonen *et al*., 1997; see, for reviews, Weigle, 2002; Hamp-Lyons, 2003).

In a seminal study, Cumming (1990) gathered think-aloud protocols and extracted no less than 28 different interpretation and judgment strategies, most of which he classified into one of three broader categories (i.e., substantive content, language use, rhetorical organization). Cumming further observed that expert raters, as compared to novices, tended to use a wider range of criteria and to integrate more effectively their interpretations and judgments of situational and textual features of the compositions. Similarly, Wolfe *et al*. (1998) demonstrated that highly proficient raters were more likely to focus on general features of an essay and to stay closer to the rating categories found in the scoring rubric than less proficient raters.

Other studies looked specifically at the cognitive and decision-making processes that are involved when trained, experienced raters holistically score examinee performance. Adopting a think-aloud approach, Vaughan (1991) identified several characteristic reading strategies or reading styles, such as the 'first-impression-dominates style', the 'two-category style' (with a focus, for example, on organization and grammar), and the 'grammar-oriented style' (with an almost exclusive focus on grammatical elements). She concluded that 'despite their similar training, different raters focus on different essay elements and perhaps have individual approaches to reading essays' (p. 120). Vaughan also observed that the use of individual reading styles was especially likely when an essay did not fit well into the scoring rubric (see also Weigle, 1994).

Taking the complexity of raters' cognitive processes into account, Milanovic *et al*. (1996) made use of multiple methods for data collection. In addition to performing group interviews, they asked individual raters to provide retrospective written reports as well as introspective verbal reports on their thought processes. Analysis of these data revealed four reading styles: 'the principled two-scan/read', 'the pragmatic two-scan/read', 'the read through', and 'the provisional mark'. In addition, Milanovic *et al*. assembled a strikingly heterogeneous list of 11 essay elements which raters focused on: *length*, *legibility*, *grammar*, *structure*, *communicative effectiveness*, *tone*, *vocabulary*, *spelling*, *content*, *task realization*, and *punctuation*. For most of these elements, the researchers were

able to show that the weight attributed to any particular element varied widely among raters. For example, some raters did not lend much weight to *vocabulary*, whereas others seemed to be strongly affected by this element.

Subsequent think-aloud studies confirmed that raters tend to display highly individual reading styles. Employing an analytic scoring method, Smith (2000) grouped six raters' verbal reports into three approaches to the task of rating, two of which were congruent with those identified in earlier research: 'the read-through-once-then-scan', 'the performance criteria-focused', and 'the first-impression-dominates'. Analysis of the data further revealed that the two raters falling into 'the first-impression-dominates' category commented on more extraneous textual features than other raters. In contrast, the two 'performance criteria-focused' raters commented on the least number of textual features, suggesting that the performance criteria statements had a heightened controlling effect on these particular raters.

In Sakyi's (2000) research on holistic scoring, four distinct styles emerged, characterized by a focus on the following: (a) errors in the text; (b) essay topic and presentation of ideas; (c) the rater's personal reaction to text; or (d) the scoring guide. On the basis of these and related findings, Sakyi built a tentative model of the holistic scoring process. In this model, content (e.g., idea development, organization, relevance) and language (grammar, mechanics, vocabulary, syntax) figured prominently as two kinds of factors affecting the formation of a general impression of an essay's quality. Two other kinds of influences specified in the model pertained to raters' personal biases/expectations and their personal monitoring processes, respectively. According to Lumley (2002; 2005), individual monitoring processes come into play when raters try to reconcile their overall impression of the text, specific, hard-to-evaluate features of the text and the rating scale descriptors. For example, in Lumley's research raters differed in the emphasis they gave to the various components of the scale descriptors (e.g., 'relevance of response' vs. 'clarity of meaning').

Building on the studies by Cumming (1990) and Sakyi (2000), Cumming *et al*. (2001; 2002) performed a fine-grained analysis of the decision-making behaviors that experienced raters of essays use. In one approach, the researchers asked raters to indicate elements they believed to characterize especially effective writing. The most frequently mentioned elements referred to rhetorical organization (e.g., introductory statements, cohesion, fulfillment of the writing task), expression of ideas (e.g., logic, argumentation, clarity), and accuracy and fluency of English grammar and vocabulary.

Cumming *et al*. (2001; 2002) also looked at the criteria that emerged from think-aloud data and identified three general categories: a self-monitoring focus (e.g., reading or rereading the essay, comparing with other essays), a rhetorical and ideational focus (e.g., assessing topic development, task completion), and a language focus (e.g., considering error frequency, lexis, syntax). The researchers concluded that 'criteria for scoring should balance equal attention to interpreting and to judging key aspects of written compositions, as well as equivalent attention to rhetoric and ideas and to language features' (p. 71).

III Context of the present study: The TestDaF writing assessment

The assessment context in which this research is situated is the writing section of the Test of German as a Foreign Language (*Test Deutsch als Fremdsprache*, TestDaF). This test is designed for foreign students applying for entry to an institution of higher education in Germany. Test tasks and items are centrally constructed and evaluated, and TestDaF examinee performance is centrally scored (for more details, see Grotjahn, 2004; Eckes *et al*., 2005; also www.testdaf.de). In spring 2001, the TestDaF was administered worldwide for the first time. In each administration, the test measures the four language skills in separate sections. Examinee performance in each section is related to one of three levels of language proficiency in the form of band descriptions; these levels (*TestDaF-Niveaustufen*, TestDaF levels, or TDNs for short) are *TDN 3*, *TDN 4*, and *TDN 5*. The TDNs are intended to cover the Council of Europe's (2001) Lower Vantage Level (B2.1) to Higher Effective Operational Proficiency (C1.2); that is, the TestDaF measures German language proficiency at an intermediate to high level. There is no differentiation among lower proficiency levels; it is just noted that the TDN 3 level has not yet been achieved (*below TDN 3*).

The writing section is a performance-based instrument, designed to assess an examinee's ability to produce a coherent and well-structured text on a given topic taken from the academic context. There is a single task, requiring two types of prose: description and argumentation. More specifically, in the first part of this section, charts, tables, or diagrams are provided along with a short introductory text, and the examinee is asked to describe the relevant information. Specific points to be dealt with are stated in the rubric. In the second part, the examinee has to consider different positions on

an aspect of the topic and write a well-structured argument. The input consists of short statements, questions, or quotes. As before, aspects to be dealt with in the argumentation are stated in the rubric.

Each examinee's performance in the writing section is scored by trained raters on a 4-category rating scale, defined by the TDN levels (i.e., *below TDN 3*, *TDN 3*, *TDN 4*, *TDN 5*). A set of nine criteria provides the basis for awarding scores (or TDN levels) to examinee performance. The criteria have been carefully developed so as to capture the gist of the underlying construct, that is, the ability to write a clearly structured and coherent text dealing with a specific topic taken from the academic context. Three of these criteria are more holistic in nature, referring to the *overall impression* upon first reading the written text, whereas the others are more of an analytic kind, referring to various aspects of *task realization* and *linguistic realization*, respectively. Each analytic criterion requires at least two readings of the text. The complete set of scoring criteria, as used in this study, is presented in the Method section.

For each criterion, raters are provided with scale descriptors specifically designed to characterize written performance at each of the four TDNs. TestDaF scale descriptors were extensively piloted and revised to make sure that they could be applied adequately by trained raters.

For example, one overall impression criterion is *fluency*; that is, the degree to which the text reads fluently. The descriptors for this criterion at the respective TDN levels are as follows: 'The text reads fluently throughout' (TDN 5), 'Readability is slightly impaired in places' (TDN 4), 'Repeated reading of parts of the text is necessary' (TDN 3), and, finally, 'On the whole, the text does not read fluently' (below TDN 3). As another example, one linguistic realization criterion is *vocabulary*; that is, the degree of differentiation and precision of the wording. In this case, the descriptors are: 'The vocabulary is varied and precise' (TDN 5), 'The vocabulary is broad, but not always precise' (TDN 4), 'The vocabulary is sufficient' (TDN 3), and, finally, 'The vocabulary is limited' (below TDN 3).

Combining the set of criteria with the set of TDNs yields 36 distinct scale descriptors. It is in the use of these descriptors that TestDaF raters are specifically trained in sample scoring sessions that take place before each TestDaF exam, focusing on the demands of the writing task employed in that exam.

In a series of studies on the psychometric quality of the TestDaF rater-mediated performance assessment system, the issue of rater variability has been closely examined (see, e.g., Eckes, 2003; 2004; 2005b). The main results of these studies can be summarized as

follows: (a) raters differed markedly in the severity with which they rated examinees, (b) raters were fairly consistent in their overall ratings, and (c) raters were substantially less consistent in relation to scoring criteria than in relation to examinees. Based on these findings, Eckes (2005a) suggested that rater behavior may be characterized by distinctive and coherent patterns of scoring tendencies, which in turn may best be captured through the concept of *rater types*. This concept and the ensuing rater type hypothesis are addressed in the following section.

IV The rater type hypothesis

The research reviewed earlier suggests several conclusions. Raters, even when they are experienced and trained in the appropriate use of the scoring rubric, differ widely in their rating behavior. More importantly, raters seem to exhibit distinct reading styles, each style characterized by rater-specific ways to focus on, and to process, essay-relevant information.

It should be noted, however, that much of the previous research relied on a single methodological approach, that is, on the qualitative analysis of think-aloud data, which, as commonly acknowledged, is limited in several respects (see, e.g., Green, 1998; Lumley, 2005; Lumley & Brown, 2005). Moreover, the majority of studies investigated raters' reading styles in a holistic scoring context, rendering the interpretation and use of an analytic scoring rubric an under-researched topic (see Smith, 2000). Finally, the number of raters observed in each study was fairly small, typically ranging from four to 10. For example, in Vaughan's (1991) often-cited research only nine raters participated and, based on their tape-recorded verbal comments while reading six essays, these raters were taken to exemplify one of five reading styles. As the author herself noted, the generalizability of the findings was questionable and clearly in need of substantiation in further research.

In the present study, I examine a relatively large sample of trained, experienced raters, look at their ways of dealing with a set of routinely-used performance criteria, and employ a quantitative classificatory approach to the identification of rater types. Rater judgements of the perceived importance of criteria as applied in operational TestDaF scoring sessions serve as the basis for building a joint classification scheme of raters and criteria. The purpose of this scheme is to systematize the between-rater differences and to yield insights into the specific ways raters diverge from, or are similar to, each other.

Specifically, the *rater type hypothesis* consists of three parts, which can be spelled out as follows:

- Experienced raters of written compositions fall into types (classes, clusters) that are characterized by attaching distinctive patterns of importance to routinely-used scoring criteria.
- Rater types are organized into a rater taxonomy, that is, into a hierarchical classification system relating raters and criteria to one another.
- The strength of the rater–criterion relationship represents the differing attentional focus of raters on scoring criteria.

Thus, according to this hypothesis raters would not only differ along some latent continuum of criteria interpretation, but would form fairly homogeneous classes that are separated from each other in terms of showing qualitatively different patterns of lending weight to criteria. Such a classification scheme, or rater taxonomy, could advance our knowledge of which kinds of raters tend to focus on which kind of criteria and, thus, may help to account for rater variability. In applied settings, this knowledge could be used for rater monitoring and rater training purposes with the goal of improving rater-mediated systems of performance assessment.

V Method

1 Participants

The total sample of participants comprised 65 raters (52 women, 13 men) who worked on the TestDaF writing section. One rater failed to provide criterion importance ratings. The raters were all experienced teachers and specialists in the field of German as a foreign language, and had been systematically trained and monitored as to compliance with scoring guidelines. Specifically, monitoring of these raters covered a three-year period (from 2002 to 2004), including 11 scoring sessions, with the number of raters per session varying between 21 and 32. Judged by rater infit and rater outfit indices, resulting from routinely run many-facet Rasch analyses, the vast majority of raters provided high-quality scoring in the sessions in which they participated (for more detail on these analyses, see Eckes, 2003; 2004; 2005b). Raters' ages ranged from 29 to 70 years ($M = 45.02$, $SD = 9.27$). Each rater was informed of the nature of the study and how the data would be used. All raters participated on a voluntary basis.

2 Instrument and procedure

Raters were presented with a questionnaire asking them to rate each criterion according to its importance for evaluating examinee performance in TestDaF's writing section. Ratings were recorded on a four-point scale, with the categories *less important*, *important*, *very important*, and *extremely important*. The raters were instructed not to think of a particular writing performance, but to indicate the weight they would *in general* attach to each criterion when working in a TestDaF scoring session.

The nine criteria to be rated for general importance corresponded in wording as closely as possible to the operational TestDaF scoring criteria, with which the raters were highly familiar. The criteria as presented to the participants in this study were as follows:

- *Fluency*: the degree to which the text can be read fluently
- *Train of thought*: the degree to which the train of thought can be followed
- *Structure*: the degree to which the text is structured
- *Completeness*: the degree to which all of the points specified in the task description are dealt with
- *Description*: the degree to which the information contained in the prompt, such as a table or diagram, is summarized
- *Argumentation*: the degree to which points of view/personal considerations are recognizable
- *Syntax*: the degree to which the text exhibits a range of cohesive elements and syntactic structures
- *Vocabulary*: the degree to which the vocabulary is varied and precise
- *Correctness*: the degree to which the text contains morphosyntactic, lexical, or orthographical errors.

In addition to the importance ratings, the questionnaire asked raters about their background training and professional experience.

3 Data analysis

a Preliminary analysis: Two-facet Rasch measurement. The purpose of the Rasch analysis was to make sure that the importance ratings provided a sound basis for the main analysis. That is, I examined whether raters differed significantly in the importance attached to criteria, whether criteria differed significantly in their perceived importance, and whether the rating scale functioned as intended, that is,

as a four-category scale in which higher categories reliably represented higher levels of criterion importance. To illustrate, if raters similarly perceived all of the criteria as very important (as might seem desirable from a theoretical point of view), or if raters used the importance rating scale in an inconsistent manner, any subsequent attempt to classify raters into types would not make much sense.

In order to address these issues, I analyzed the importance rating data by means of the Rasch computer program FACETS (Version 3.59; Linacre, 2005). FACETS used these ratings to estimate measures of the overall importance individual raters attached to the set of criteria, as well as measures of the importance of individual criteria as perceived by the raters. In addition, the program computed several indices of the effectiveness of the rating scale categories.

It should be kept in mind, though, that the major purpose of the FACETS analysis in the present research was to ascertain the degree of differentiation among raters' perceptions of criterion importance and to examine the functioning of the importance rating scale. Since the raters did not evaluate real examinee performance, but rated the general importance of familiar scoring criteria (forming a rater × criterion factorial design), some FACETS-based methods of checking the psychometric quality of ratings were not applicable or took on different meanings here (for detailed discussions of such methods, see Engelhard, 2002; Myford and Wolfe, 2003; 2004). In particular, if the rating data were to bear out the rater type hypothesis, rater fit statistics needed to indicate substantial degrees of deviations from model expectations. That is, there needed to be cases of rater overfit, showing that the respective raters perceived many criteria as similarly high in importance. Conversely, there also needed to be cases of rater misfit showing that a different set of raters perceived specific criteria as highly important and, at the same time, other criteria as much less important. To pin down which raters would go with which criteria was the purpose of the analysis outlined next.

b Main analysis: Two-mode clustering: In the main analysis I employed a two-mode clustering approach to test the rater type hypothesis. Since this approach is not as widely known in the field of language testing and assessment as is its measurement counterpart (i.e., the many-facet Rasch measurement model), a short outline of the basic two-mode clustering model and the specific technique used seems in order.

The general goal of two-mode cluster analysis is to construct a common categorical representation of two different sets of elements,

such as a set of raters and a set of scoring criteria.[1] Two-mode clustering is especially indicated when the sets involved have equivalent theoretical or empirical status; that is, when neither set has priority over the other one, as is the case in this research. Use of two-mode clustering further supports the interpretation of clusters identified through cross-referencing between the modes, provides a detailed look at the structure of inter-mode relationships, and allows a direct analysis of non-symmetric proximity data (see Eckes and Orlik, 1993; Eckes, 1996; Everitt *et al*., 2001). It is in these respects, then, that two-mode clustering is superior to a traditional clustering approach, in which one mode (e.g., raters) is clustered separately from the other (e.g., criteria), and the researcher is left with the unwieldy task of relating the resulting clustering solutions to each other.

Several models and algorithms have been developed to perform a two-mode cluster analysis (for a comprehensive review, see Van Mechelen, Bock, & De Boeck, 2004). In this study, the error-variance technique introduced by Eckes and Orlik (1991; 1993) was employed. The error-variance technique simultaneously represents raters and scoring criteria by means of a hierarchical clustering system (for in-depth discussions of this technique, see Mirkin *et al*., 1995; Castillo and Trejos, 2000; Everitt *et al*., 2001: 154–161).[1]

At each consecutive step of the algorithm, a particular subset of the first mode is merged with a particular subset of the second such that the increase in an internal heterogeneity measure of the resulting two-mode cluster is at a minimum. This measure, the *mean squared deviation index* (MSD index), takes into account both the variance within each candidate cluster and its 'centroid effect', defined as the squared deviation of its mean from the maximum entry in the input matrix. That is, a two-mode cluster is said to have low internal heterogeneity to the extent that its elements show strong inter-mode relations with as small a variance of corresponding numerical values as possible.

A decision on the number of clusters can be reached by using a stopping rule analogous to that in traditional one-mode hierarchical clustering: the step-size criterion *MSD increase*. That is, a marked

[1] The terms *mode* (as used in two- or higher-mode clustering) and *facet* (as used in many-facet Rasch measurement) have identical meanings in principle; that is, both refer to a particular set of entities considered for analysis. The differences that exist between them have etymological roots: Whereas the mode concept goes back to Tucker's (1964) notion of three-mode factor analysis, the facet concept was originally introduced by Guttman (1959) in his work on facet theory.

increase in fusion values from one level to the next is considered indicative of the formation of a relatively heterogeneous cluster. In addition, a criterion related to the heterogeneity index may be used. This criterion, the *centroid effect ratio* (CER), takes on values between 0% and 100%, with larger values indicating clusters having higher inter-mode cohesion.

As a final option, the algorithm allows the construction of an *overlapping* clustering solution in which particular raters and/or criteria may be assigned to more than one cluster at a given hierarchical level. Whereas the initial clustering solution comprises clusters that are disjoint or mutually exclusive at a particular level of the hierarchy (i.e., none of the elements at the same level of abstraction can belong to more than one cluster), it would seem much more appropriate to let raters and criteria already belonging to a particular cluster be placed into as many other clusters as reasonable given the two-mode similarity structure of the input data.

The basic procedure is as follows: having decided on the number of (disjoint) clusters, each row and column element of the input matrix not already belonging to a given cluster is considered in turn as a possible candidate for joining it. An element is added if the corresponding fusion criterion does not show a marked increase and/or the CER index of the resulting cluster remains sufficiently high. In fact, in the analyses reported below, a combination of both criteria was employed, with the critical CER value set as high as 95% to ensure that only very closely associated elements were added and to prevent the overlapping solution from becoming an unwieldy representation of the data.

In this research, the cluster analysis proceeded in three steps. In the first, preparatory step the rating data were arranged in a matrix with 64 raters (row elements) and nine scoring criteria (column elements). Each column was duplicated, and then the duplicated entries were rescored (i.e., rating category ‘4’ was changed to ‘1’, category ‘3’ changed to ‘2’, etc.). This was to make sure that criteria perceived as *extremely important* could be clustered separately from criteria perceived as *less important*. Each reflected criterion was marked with a minus sign for the purpose of identification. In the second step, the augmented 64×18 data matrix was subjected to the hierarchical error-variance algorithm. Finally, after deciding on the number of clusters, an overlapping clustering solution was constructed.

VI Results

1 Two-facet Rasch analysis

Overall data–model fit was satisfactory, as judged by the percentage of standardized residuals exceeding pre-specified levels (see Linacre, 2005). Thus, only 3.12% of the (absolute) standardized residuals were equal to or greater than 2, and only 0.17% of the residuals were equal to or greater than 3 (the critical percentage values are conventionally set at 5% and 1%, respectively).

Figure 1 displays the variable map representing the calibrations of raters (identified by numbers only), criteria, and the four-category rating scale as raters used it to rate each criterion. Table 1 provides summary statistics from the FACETS analysis.

As can be seen, the variability across rater measures as well as across criterion measures was substantial. This was consistently confirmed by three kinds of separation statistics: (a) for both facets, the fixed chi-square value was highly significant, leading to a rejection of the null hypothesis that all raters shared the same view of the importance of scoring criteria, and also the null hypothesis that all criteria were viewed as similarly important; (b) the separation index showed that there were around three distinct strata of raters and between five and six distinct strata of criteria; and, finally, (c) the separation reliability indices demonstrated that raters were well-differentiated in terms of their perception of criterion importance and that criteria were extremely well differentiated in terms of their perceived importance.

Further analyses revealed that the rating scale functioned as intended (see Table 2). The distribution of ratings (as indicated by the relative frequencies) was spread out across the four rating scale

Table 1 Summary statistics for the many-facet Rasch analysis of criterion importance ratings

Statistics	Raters	Criteria
Mean measure	0.03	0.00
Mean *SE*	0.46	0.17
Chi-square	255.0*	140.2*
df	63	8
Separation index	3.20	5.65
Separation reliability	.82	.99

Note: $*p < .01$.

Logit	Rater	Criterion	Scale
	Attaching High Importance	*Highly Important*	
3			(4)
	06 29		
	16 32		
2	41		
	01		-----
	25 31 34		
	02	train of thought	
1	04 62		
	14 18 35 36 37 40 45	argumentation	3
	03 38 60	fluency	
	49 56		
0	11 12 43 48 53		-----
	09 39 59 64	completeness vocab. description	
	05 13 23		
	08 20 22 26 27 50 52 55 61 63	syntax structure	
	19 21 44 54 57 58		
	47		2
-1	07 17 24 46		
	10 15 33	correctness	
	28		

	51		
-2			
	30 42		
-3			(1)
	Attaching Low Importance	*Less Important*	

Figure 1 Variable map from the two-facet Rasch analysis of the importance rating data. The horizontal dashed lines in the fourth column indicate the category threshold measures for the four-point rating scale

Table 2 Functioning of the criterion importance rating scale

Category	Freq. %	Threshold	Average measure	Outfit
Less important	22%	–	−1.27	0.9
Important	26%	−0.99	−0.34	1.5
Very important	31%	−0.22	0.34	1.0
Extremely important	21%	1.21	1.36	0.9

Note: Threshold is the category threshold measure (in logits). Average measure is the average criterion importance measure (in logits) per rating category. Outfit is a mean-square fit statistic.

categories, and the category thresholds advanced monotonically with rating scale categories, as did the average rater importance measures computed per category. Moreover, the outfit mean-square values for the rating scale categories were reasonably close to the expected value of 1 (see Linacre, 2004).

As summarized in Table 3, rater infit and outfit indices showed a considerable degree of variation. Nearly half the infit values and exactly half the outfit values fell outside the narrow (0.7/1.3) fit range; still a fifth of the infit values and a quarter of the outfit values fell outside the wider (0.5/1.5) fit range. Thus, in line with the rater type hypothesis, raters produced heterogeneous patterns of importance ratings.

The Rasch statistics for the criterion measurement are set out in Table 4. The mean-square fit statistics were near the expected value of 1, with the exception of the outfit value for *train of thought*; however, according to Linacre (2002), outfit values falling outside the 0.5/1.5 fit range are less of a threat to measurement than exceedingly large (or small) infit values. Overall, then, these findings supported the assumption of unidimensionality within the set of scoring criteria (see also Smith, 2004).

In sum, the two-facet Rasch analysis demonstrated that the raters differed significantly in the importance they attached to the scoring criteria, and the rating scale categories were functioning effectively. These findings provided the basis for the main analysis, that is, for the two-mode clustering that aimed at the identification of rater types.

2 Two-mode cluster analysis

Using both the step-size criterion (i.e., a sharp increase in the MSD index) and the CER index (with a lower limit of 90%) as stopping rules, six two-mode clusters seemed to provide the best trade-off

Table 3 Absolute and relative frequencies of rater fit statistics

Fit range	Infit		Outfit	
	n	%	n	%
Narrow				
fit < 0.70 (overfit)	17	26.6	20	31.2
0.70 ≤ fit ≤ 1.30	35	54.7	32	50.0
fit > 1.30 (misfit)	12	18.7	12	18.7
Wide				
fit < 0.50 (overfit)	8	12.5	11	17.2
0.50 ≤ fit ≤ 1.50	50	78.1	44	68.7
fit > 1.50 (misfit)	6	9.4	9	14.1

Note: Infit and outfit are mean-square fit statistics. Total number of raters = 64.

Table 4 Measures and fit statistics of scoring criteria

Criterion	Measure	*SE*	Infit	Outfit
Fluency	0.55	0.16	0.99	1.05
Train of thought	1.16	0.18	0.98	1.73
Structure	−0.53	0.16	1.12	1.11
Completeness	−0.07	0.16	1.21	1.12
Description	−0.09	0.16	0.94	0.90
Argumentation	0.80	0.17	1.28	1.20
Syntax	−0.50	0.16	0.70	0.66
Vocabulary	−0.17	0.16	0.70	0.66
Correctness	−1.15	0.18	0.93	0.95

Note: Measure = Importance in logits. *SE* = Standard error. Infit and outfit are mean-square fit statistics.

between an economical representation of the rating data and interpretability of the clustering solution. All but one of these clusters were affected by the additional overlapping clustering procedure.

Table 5 shows the cluster statistics for the non-overlapping solution. For ease of reference, clusters are labeled A through F. As is evident from this table, Cluster B forms a singleton containing only one rater who attached extremely high importance to just one criterion, namely *correctness*. This cluster is clearly reminiscent of the grammar-oriented reading style identified by Vaughan (1991). The other clusters each represent more varied, yet distinct patterns of relationships between raters and criteria. Aside from Cluster B, the most homogeneous clusters in terms of the MSD and CER indices are Clusters C, E, and F. Somewhat less homogeneity holds for Clusters A and D,

Table 5 Cluster statistics for the non-overlapping six-cluster solution

Cluster	Number of raters	Number of criteria	*M*	*SD*	MSD	CER
A	17	5	3.16	0.85	1.42	93
B	1	1	4.00	0.00	0.00	100
C	12	2	3.58	0.57	0.50	98
D	23	6	3.28	0.77	1.11	95
E	5	2	3.60	0.49	0.40	98
F	6	2	3.50	0.65	0.67	97

Note: MSD = Mean squared deviation (cluster heterogeneity index; minimum MSD = 0, higher values indicate greater heterogeneity; see Eckes and Orlik, 1993). CER = Centroid effect ratio (cluster cohesion index; minimum CER = 0, maximum CER = 100, higher values indicate greater cohesion; see Eckes & Orlik, 1993).

but this is easily accounted for by the fact that these two clusters contain by far the largest numbers of raters and criteria. It should be remembered that extremely important criteria were clustered separately from less important criteria; hence the total number of criteria considered was 18.

Figure 2 simultaneously portrays the non-overlapping and the overlapping solutions. The cluster elements that were added by the overlapping clustering procedure are indicated by dashed rectangles. Criteria that are specific to a given cluster (i.e., not shared with any other cluster) are in boldface. For ease of presentation, only the highest fusion levels of the resulting hierarchical classification system are shown.

Cluster A contains raters with a strong focus on criteria referring to linguistic realization (i.e., *vocabulary*, *syntax*) and to task realization (i.e., *argumentation*, *completeness*). From the overall impression category only *train of thought* is added. As indicated by the minus sign for *structure*, raters belonging to Cluster A put much less weight on the overall organization of a text. Since *vocabulary* and *syntax* are the only criteria that are specific to this type, Cluster A may be called the *Syntax Cluster* or *Syntax Type* for short.

Quite a different focus is exhibited by Cluster B raters. These raters focus on the complete set of criteria from the task realization category, as well as on *correctness* as the sole criterion from the linguistic realization category. Note also that *correctness* is specific to this cluster, hence the corresponding rater type may be given the name *Correctness Type*.

For Cluster C raters, the most important criteria belong to the global impression and task realization categories. The *structure*

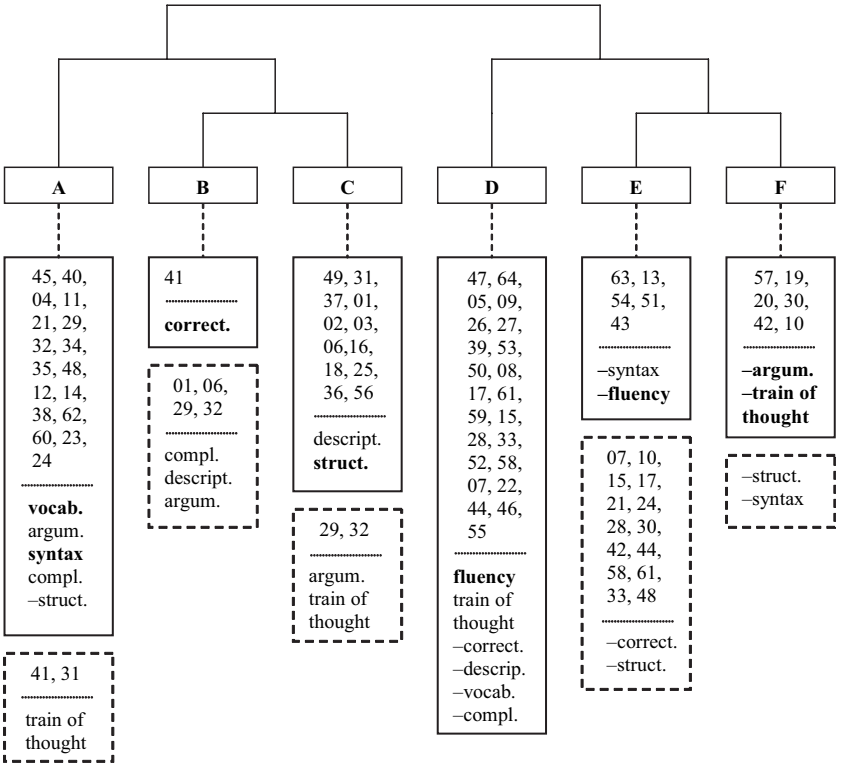

Figure 2 Two-mode clustering solution from the error-variance analysis of the importance rating data. Cluster elements added in the process of constructing an overlapping solution are indicated by dashed rectangles. Raters (identified by numbers) are shown together with criteria they viewed either as highly important or as less important (less important criteria are marked by a minus sign)

criterion forms one of these, thus epitomizing a view that clearly deviates from Cluster A raters' perceptions. Moreover, this criterion is not shared with any other cluster, suggesting the name *Structure Type* for Cluster C.

Raters belonging to Clusters D through F tend to attach much less importance to criteria overall than raters of Clusters A through C. Particularly intriguing are the differences between the two largest clusters, A and D: the only criterion on which raters from both clusters agree is *train of thought*. All the other criteria are weighted differently, with *completeness* and *vocabulary* standing out as receiving contrasting ratings. Moreover, there is just one criterion (i.e., *fluency*) that is specific to Cluster D. Thus, this rater type can be termed the *Fluency Type*.

Though there are some commonalities between Clusters D and E, the major difference between the two is that Cluster E comprises raters who specifically put much *less* weight on the criterion of *fluency* (the *Non-fluency Type* for short).

Finally, Cluster F contrasts most clearly with Cluster C, the *Structure Cluster*. Quite obviously, Cluster F raters view both *argumentation* and *train of thought* as much less important, and they do so in ways that are specific to this cluster (the *Non-argumentation Type*).

The patterns of commonalities and differences between clusters are summarized in Table 6. In this table, 'Hi' is shorthand for raters of a given cluster perceiving a particular criterion as *extremely important* (i.e., most or all raters in the cluster marked '4' on the rating scale), 'Mo' refers to viewing a particular criterion as *moderately important* (i.e., most or all raters in the cluster marked '2' or '3' on the rating scale), and 'Lo' stands for viewing a particular criterion as *less important* (i.e., most or all raters in the cluster marked '1' on the rating scale).

Portrayed this way, it becomes even more evident that Clusters A (*Syntax Type*) through C (*Structure Type*) are much closer to the situation where *all* criteria attain maximum importance, than are Clusters D (*Fluency Type*) through F (*Non-argumentation Type*). Yet, even Clusters A through C show points of divergence from this reference line, and they do so in different ways; that is, each cluster has its own distinct profile of perceived scoring criterion importance. What is more, some cluster, or rater type, profiles are in sharp contrast with one another, which means that raters of different types showed markedly different scoring foci (compare again the *Syntax Type* with the *Fluency Type*).

The fact that Cluster E and Cluster F raters did not perceive any of the criteria as *extremely important* raises the question as to whether rater training failed to succeed in establishing a clear understanding of the weight each criterion should generally have in TestDaF essay scoring. It may also mean that these raters use criteria that are different from those they are supposed to use. In any case, this outcome from the clustering analysis is a clear indication of the need to probe more deeply into raters' cognitions of scoring criteria, possibly through interviews and other individual-centered approaches. Eventually, these more qualitative routes to charting raters' cognitive maps of the scoring criteria, along with the rater type classification system, can inform the implementation of specific rater monitoring and retraining procedures.

Aside from these possible practical implications of the present results, the overall high degree of cluster profile distinctiveness

Table 6 Cluster profiles of perceived scoring criterion importance

Cluster	Global impression			Task realization			Linguistic realization		
	Fluency	Train of thought	Structure	Completeness	Description	Argumentation	Syntax	Vocabulary	Correctness
A (*Syntax*)	Mo	Hi	Lo	Hi	Mo	Hi	Hi	Hi	Mo
B (*Correctness*)	Mo	Mo	Mo	Hi	Hi	Hi	Mo	Mo	Hi
C (*Structure*)	Mo	Hi	Hi	Mo	Hi	Hi	Mo	Mo	Mo
D (*Fluency*)	Hi	Hi	Mo	Lo	Lo	Mo	Mo	Lo	Lo
E (*Non-fluency*)	Lo	Mo	Lo	Mo	Mo	Mo	Lo	Mo	Lo
F *(Non-argumentation)*	Mo	Lo	Lo	Mo	Mo	Lo	Lo	Mo	Mo

Note: 'Hi' indicates extremely *high importance*. 'Mo' indicates *moderately high importance*. 'Lo' indicates *low importance*.

suggests that any attempt to predict differences in perceived importance of scoring criteria on the basis of rater background variables would need to take rater types into account. The next section addresses this issue.

3 Correlations with rater background variables

Aside from the expert–novice kind of comparison studies cited earlier, there has been surprisingly little empirical work on the influence of rater background variables on the interpretation and use of scoring criteria in writing performance assessment. One notable exception is Myford *et al.*'s (1996) work in the context of the Test of Written English (TWE). These researchers studied various rater characteristics, such as the number of languages read/spoken or the number of TWE readings, but they did not find strong, consistent correlations with measures of rating performance. One explanation for this failure may be that their sample of raters was composed of distinct rater types, each with a specific pattern of predictive relationships canceling one another out when all raters were considered as one large group.

To gain some insight into this issue I computed cluster-specific correlations of various rater background variables with (a) rater measures of perceived criterion importance (from the present many-facet Rasch analysis) and (b) rater severity measures (aggregated over TestDaF scoring sessions in which raters had participated). The prediction was that rater types would show marked differences in the pattern of correlations across background variables, providing evidence for the claim that variability in rating behavior is differentially associated with raters' personal and professional characteristics. Table 7 presents the correlations with perceived criterion importance for Clusters A, C, and D, as well as the total sample of raters. Only these three clusters had sufficiently large numbers of raters to be considered separately.

Although at this stage of the research the correlation analysis was merely exploratory in nature, some remarkable results showed up. First of all, in the total sample of raters the correlations were consistently low and failed to reach the level of significance. Quite in contrast, each of the three clusters was characterized by a different pattern of correlations, some of which were statistically significant and fairly strong.

Thus, Cluster A, the *Syntax Cluster*, was the only cluster yielding a significantly negative correlation between rater age and rater-perceived importance measures; that is, older raters tended to view

Table 7 Cluster-specific correlations of rater-perceived importance measures with rater background variables

Cluster	Number of raters	Age	Number of FLs spoken	Years abroad	Years as GFL teacher	Years as GFL examiner	Number of scoring sessions	Mean rater severity
A (*Syntax*)	17	−0.55*	0.51*	−0.04	−0.07	−0.51*	0.01	0.42
C (*Structure*)	12	0.26	0.34	0.64*	0.14	0.18	0.24	−0.62*
D *(Fluency)*	23	0.28	−0.42*	−0.11	0.27	0.12	−0.46*	0.20
All (A–F)	64	−0.07	0.22	0.11	−0.05	−0.06	0.06	0.08

Note: Clusters B (*Correctness*), E (*Non-fluency*), and F (*Non-argumentation*) were not considered because the respective number of raters was too small. Mean rater severity was computed as the mean percentage rank of a particular rater's severity collapsed across all scoring sessions in which he or she participated. * $p < .05$.

the criteria as being of lower importance overall than did younger raters. Conversely, rater-perceived importance measures were significantly correlated with the number of years raters spent abroad for Cluster C, the *Structure Cluster*, but not for any of the other clusters. Moreover, this cluster yielded the only significant correlation with rater severity measures, and it was a negative one; that is, raters seeing the criteria as more important overall tended to be more lenient when awarding scores to examinees. Finally, Cluster D, the *Fluency Cluster*, showed negative correlations between perceived criterion importance and (a) the number of foreign languages spoken, which is in direct contrast to Cluster A, and (b) the number of TestDaF scoring sessions.

In the case of the correlations between rater background variables and rater severity measures (Table 8), only one significant cluster-specific result emerged: the correlation with rater age for Cluster D (*Fluency Type*). There was also a significant correlation for the whole sample with the number of years as a teacher in the field of German as a foreign language (GFL). The latter correlation remained positive and at a similar level across clusters.

VII Summary and discussion

In this research, I examined the rater variability issue from a classificatory perspective. The rater type hypothesis advanced here states

Table 8 Cluster-specific correlations of rater severity measures with rater background variables

Cluster	Number of raters	Age	Number of FLs spoken	Years abroad	Years as GFL teacher	Years as GFL examiner	Number of scoring sessions
A (*Syntax*)	17	−0.05	−0.17	0.10	0.38	−0.14	0.25
C (*Structure*)	12	−0.01	−0.36	−0.35	0.32	0.06	−0.03
D *(Fluency)*	23	0.49*	0.01	−0.23	0.31	0.35	0.18
All (A–F)	64	0.10	0.12	−0.22	0.30*	0.22	0.22

Note: Clusters B (*Correctness*), E (*Non-fluency*), and F *(Non-argumentation)* were not considered because the respective number of raters was too small. * $p < .05$.

that experienced raters operating on a large-scale performance assessment instrument fall into types (classes or clusters), each representing a distinct pattern of attaching importance to routinely-used scoring criteria. According to this hypothesis, each rater type is characterized by close interconnections between a particular subset of raters and a particular subset of criteria. Raters of a given type view some criteria as highly important and other criteria as less important in scoring examinees' writing performance. Put differently, each rater type is assumed to be characterized by a specific scoring focus or scoring profile. The present results strongly support this hypothesis.

Based first on a many-facet Rasch measurement approach to the analysis of the importance rating data, results show the following.

- Raters differed significantly in their views of the general importance of nine routinely used and well-specified scoring criteria. In addition, the pronounced variation in rater fit statistics suggests that raters' perceptions of the scoring criteria lacked common ground.
- Criteria differed significantly in their perceived general importance for scoring examinee performance. At the same time, criterion fit statistics provide evidence for the unidimensionality of criterion importance ratings.
- The four-point scale along which raters judged the perceived importance of criteria functioned effectively.

Taken together, these findings substantiate the claim that there is a significant degree of rater variability in the importance rating data,

thus setting the stage for the main part of this study – the construction of a joint rater × criterion classification system. The two-mode clustering analysis employed here reveals the following:

- Six rater types emerged, each of which was characterized by a distinct scoring profile; that is, each rater type represented a markedly different point of view concerning criterion importance. Four of these types were defined by criteria standing out as extremely important: the *Syntax Type*, the *Correctness Type*, the *Structure Type*, and the *Fluency Type*. The remaining two types were defined by criteria to which raters gave specifically less weight: the *Non-fluency Type* and the *Non-argumentation Type*.
- Further in line with the rater type hypothesis, none of the types divided their attention evenly across the complete set of criteria; rather, each type seemed to attend to a different subset of criteria.
- Compared across rater types, there was evidence of complementary scoring foci. For example, the focus of the *Fluency Type* was exclusively on overall impression criteria, whereas the *Syntax Type* focused on criteria primarily referring to task realization and to linguistic realization.

Furthermore, findings from the correlation analysis confirm that rater background variables were differentially associated with type-specific rater measures of perceived criterion importance – quite in contrast to the total rater sample, in which the between-cluster profile differences seemed to level out. Thus, for example, perceived criterion importance was significantly *positively* correlated with the number of foreign languages spoken in the *Syntax Type*, yet significantly *negatively* correlated with this background variable in the *Fluency Type*. It appears, then, that explanatory studies of rater variability need to take type-specific correlation patterns into account.

Quite obviously, the present results have implications for rater monitoring, rater training, and further research into rater variability. The study of rater types can inform rater monitoring processes by highlighting stability and change in type-specific scoring foci and, even more importantly, shed light on the validity of the assessment procedure as a whole. Thus, when a given examinee performance is scored by only a small number of trained raters arbitrarily selected from a larger group of raters, as is the case usually in large-scale assessments, and these raters differ from others in their scoring profile, an ‘element of chance’ (to borrow a term from Edgeworth, 1890) is added to the procedure over and above differences in rater severity or other well-documented judgmental tendencies.

Moreover, raters sharing a specific scoring profile are expected to be in close agreement with one another, which typically would be evidenced by high coefficients of interrater reliability. However, the validity of the scores these raters award to examinees would seem questionable. In that instance, high interrater reliability could simply be due to these raters' type-specific points of view regarding the weight of scoring criteria, which actually may capture only a small part of the construct being assessed (for a similar point, see Shohamy, 1995; Reed & Cohen, 2001).

Rater training can help to resolve these problems. But how much raters' scoring profiles will be influenced through specific training is currently an unaddressed issue. In any case, if distinct rater types can be identified in a given assessment context, rater training could concentrate on redirecting the attention of particular rater types to criteria not captured within their respective scoring profiles, thus contributing to a more balanced use of the criteria deemed relevant in the assessment.

The conceptual distinction between *behavior-driven* (or bottom-up) rater training and *schema-driven* (or top-down) rater training (Pulakos, 1986; Lievens, 2001) could inform the design and implementation of training procedures directed at raters belonging to different rater types. Behavior-driven training divides the rating process into three phases that are strictly distinguished from each other: behavioral observation, classification, and evaluation. Raters are taught to proceed from one phase to the next only when the previous one is finished. In other words, raters are trained to classify pieces of factual information into discrete categories to form a judgment. Use of an analytic rubric fits into this approach to rater training. Conversely, in schema-driven training, raters are instructed to process information in a top-down manner, with no strict separation between observation and evaluation. The raters are taught to use a mental schema or cognitive prototype that represents effective performance at a particular level of competence. That is, they are encouraged to 'scan' the performance for schema-relevant incidents and to form online evaluations, as exemplified by training in the use of holistic scoring rubrics. In light of the present findings, it seems advisable to specifically retrain raters who display a narrow scoring focus using a bottom-up approach, instructing them to attend to identifiable features of language performance in a piecemeal fashion. In assessment contexts where holistic scoring is desired, rater training may proceed to a top-down approach or start anew with instilling into raters a mental schema of language competence.

With respect to future research, one issue concerns the extent to which the basic clustering approach to the study of rater types introduced here generalizes to assessment settings with different rating tasks, different sets of criteria, and different language skills (see Eckes, 2006 for an application to speaking data). Another issue refers to the kind of cognitive processes that possibly distinguish between raters belonging to different types – an issue that could be examined through some combination of the present clustering methodology with more process-oriented approaches like verbal protocol analysis (Green, 1998; Lumley, 2005) or quantitative techniques based on signal detection theory (DeCarlo, 2005).

In the present study, raters were asked to rate the perceived importance of scoring criteria in general, that is, abstracted from real examinee performance. Thus, one line of research that suggests itself is to apply the rater classification approach to operational rating behavior. It remains to be seen which kind of rater types would emerge when raters provided importance ratings with particular examples of writing performance in mind, and how these types would relate to the types identified in the rater cognition study discussed here.

Even more to the point, future research needs to address more closely the question of how raters' self-reported ways of perceiving scoring criteria relate to operational rater behavior. The present rater type analysis has already demonstrated that at least for some raters the perceived importance of criteria was strongly (negatively) related to rater severity collapsed across a number of operational scoring sessions. Yet, what precisely caused this correlation to become negative in one rater type (the *Structure Type*) and positive, though only marginally significant, in a second one (the *Syntax Type*), is currently unclear. Moreover, another characteristic of operational rater behavior that future research may look at is the consistency with which each rater uses particular scoring criteria. For example, one could hypothesize that the bias sizes for the interaction between raters and criteria (as revealed in many-facet Rasch analyses) are dependent on the perceived importance of the criteria, with smaller bias sizes associated with criteria viewed as more important, and vice versa. These correlation patterns would of course be predicted to vary significantly between rater types.

A further issue that has some potential for future research concerns the level of examinee proficiency vis-à-vis the perceived importance of scoring criteria. Research has shown that examinee proficiency level has an effect on the performance features that raters focus on (see, e.g., Milanovic *et al*. 1996; Pollitt & Murray, 1996). Hence, it seems worth studying whether scoring profiles generalize across levels of proficiency or change systematically from one level to the next.

Finally, further theoretical work on the antecedents and consequences of rater variability, rater background variables, including personal background, professional training, and work experience, could eventually be incorporated into structural models that would lend themselves to confirmatory analysis (see Pula & Huot, 1993). Such models would not only help to find out more precisely which rater background variables relate to measures of rater variability, but also help to explain why these relations exist in the first place.

VIII Conclusion

Explaining rater variability has been one of the biggest challenges for language assessment researchers to date. The classification approach advanced in this paper promises to help account for the structural complexity of some of the variables involved. According to Cumming *et al*. (2001: 72), 'Scoring written essays is a fundamentally interpretive and judgmental activity, based on prevailing norms of educational practice as well as individuals' past experiences.' The basic indeterminacy of essay scoring (Lumley, 2005) may be moderated, at least to some extent, by studying rater types, each imbued with its own distinct kind of regularity.

Acknowledgements

Portions of this research were presented at the 2nd Annual Conference of the European Association for Language Testing and Assessment (EALTA), Voss, Norway, June 2005, and at the 27th Language Testing Research Colloquium (LTRC), Ottawa, Canada, July 2005. I would like to thank three anonymous *Language Testing* reviewers for their valuable comments and suggestions on earlier versions of this article.

IX References

Bachman, L. F., Lynch, B. K. & **Mason, M.** (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12, 238–57.

Bachman, L. F. & **Palmer, A. S.** (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, 2, 49–58.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, 25, 587–603.

Castillo, W. & **Trejos, J.** (2000). Recurrence properties in two-mode hierarchical clustering. In R. Decker & W. Gaul, editors, *Classification and information processing at the turn of the millennium* (pp. 68–73). Berlin: Springer-Verlag.

Congdon, P. J. & **McQueen, J.** (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37, 163–178.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7, 31–51.

Cumming, A., Kantor, R. & **Powers, D. E.** (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series, MS-22). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R. & **Powers, D. E.** (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67–96.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, 42, 53–76.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, 5, 7–29.

Eckes, T. (1996). Recent developments in multimode clustering. In W. Gaul & D. Pfeifer, editors, *From data to knowledge: Theoretical and practical aspects of classification, data analysis, and knowledge organization* (pp. 151–158). Berlin: Springer-Verlag.

Eckes, T. (2003). Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse [Assuring the quality of the TestDaF: Concepts, methods, results]. *Fremdsprachen und Hochschule*, 69, 43–68.

Eckes, T. (2004). Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im "Test Deutsch als Fremdsprache" (TestDaF) [Rater agreement and rater severity: A many-facet Rasch analysis of performance assessments in the "Test Deutsch als Fremdsprache" (TestDaF)]. *Diagnostica*, 50, 65–77.

Eckes, T. (2005a). Assuring the quality of TestDaF examinations. Paper presented at the 2nd International Conference of the Association of Language Testers in Europe (ALTE), Berlin, Germany.

Eckes, T. (2005b). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2, 197–221.

Eckes, T. (2006). Raters' perceptions of scoring criteria in writing and speaking performance assessments. Paper presented at the 28th Language Testing Research Colloquium (LTRC), Melbourne, Australia.

Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & **Tsagari, C.** (2005). Progress and problems in reforming public language

examinations in Europe: Cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France, and Germany. *Language Testing*, 22, 355–377.

Eckes, T. & **Orlik, P.** (1991). An agglomerative method for two-mode hierarchical clustering. In H.-H. Bock & P. Ihm, editors, *Classification, data analysis, and knowledge organization: Models and methods with applications* (pp. 3–8). Berlin: Springer-Verlag.

Eckes, T. & **Orlik, P.** (1993). An error variance approach to two-mode hierarchical clustering. *Journal of Classification*, 10, 51–74.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, 53, 460–475, 644–663.

Elder, C., Knoch, U., Barkhuizen, G. & **von Randow, J.** (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, 2, 175–196.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31, 93–112.

—— (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna, editors, *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Lawrence Erlbaum.

Engelhard, G., Jr. & **Myford, C. M.** (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model* (College Board Research Report No. 2003–1). New York: College Entrance Examination Board.

Everitt, B. S., Landau, S. & **Leese, M.** (2001). *Cluster analysis* (4th ed.). London: Arnold.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluation. *Journal of Educational Psychology*, 71, 328–338.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Grotjahn, R. (2004). TestDaF: Theoretical basis and empirical research. In M. Milanovic & C. J. Weir, editors, *European language testing in a global context: Proceedings of the ALTE Barcelona Conference July 2001* (pp. 189–203). Cambridge: Cambridge University Press.

Guttman, L. (1959). A structural theory for intergroup beliefs and action. *American Sociological Review*, 24, 318–328.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll, editor, *Exploring the dynamics of second language writing* (pp. 162–189). Cambridge: Cambridge University Press.

Hinkel, E. (1994). Native and nonnative speakers' pragmatic interpretations of English texts. *TESOL Quarterly*, 28, 353–376.

Hoyt, W. T. & **Kerns, M.-D.** (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, 4, 403–424.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot, editors, *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton Press.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, 86, 255–264.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith & R. M. Smith, editors, *Introduction to Rasch measurement* (pp. 258–278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2005). *A user's guide to FACETS: Rasch-model computer programs* [Software manual]. Chicago, IL: Winsteps.com.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19, 246–276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Lang.

Lumley, T. & **Brown, A.** (2005). Research methods in language testing. In E. Hinkel, editor, *Handbook of research in second language teaching and learning* (pp. 833–855). Mahwah, NJ: Lawrence Erlbaum.

Lumley, T. & **McNamara, T. F.** (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12, 54–71.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, 7, 52–75.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Milanovic, M., Saville, N. & **Shuhong, S.** (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville, editors, *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (*Studies in language testing*, Vol. 3, pp. 92–114). Cambridge: Cambridge University Press.

Mirkin, B., Arabie, P. & **Hubert, L. J.** (1995). Additive two-mode clustering: The error-variance approach revisited. *Journal of Classification*, 12, 243–263.

Myford, C. M., Marr, D. B. & **Linacre, J. M.** (1996). *Reader calibration and its potential role in equating for the Test of Written English* (TOEFL Research Report No. 95–40). Princeton, NJ: Educational Testing Service.

Myford, C. M. & **Wolfe, E. W.** (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386–422.

Myford, C. M. & **Wolfe, E. W.** (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189–227.

Pollitt, A. & **Murray, N. L.** (1996). What raters *really* pay attention to. In M. Milanovic & N. Saville, editors, *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (*Studies in language testing*, Vol. 3, pp. 74–91). Cambridge: Cambridge University Press.

Pula, J. J. & **Huot, B. A.** (1993). A model of background influences on holistic raters. In M. M. Williamson and B. A. Huot, editors, *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. (pp. 237–265). Cresskill, NJ: Hampton Press.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes* 38, 76–91.

Reed, D. J. and **Cohen, A. D.** (2001). Revisiting raters and ratings in oral language assessment. In C. Elder *et al*., editors, *Experimenting with uncertainty: Essays in honour of Alan Davie* (pp. 82–96). Cambridge: Cambridge University Press.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In J. J. Kunnan, editor, *Fairness and validation in language assessment* (pp. 129–152). Cambridge: Cambridge University Press.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing,* 22, 1–30.

Schoonen, R., Vergeer, M. & **Eiting, M.** (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, 14, 157–184.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*,15, 188–211.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley, editor, *Studies in immigrant English language assessment*, Vol. 1, (pp. 159–189). Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith, editors, *Introduction to Rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.

Tucker, L. R. (1964). The extension of factor analysis to three-dimensional matrices. In N. Frederiksen & H. Gulliksen, editors, *Contributions to mathematical psychology* (pp. 109–127). New York: Holt, Rinehart and Winston.

Van Mechelen, I., Bock, H.-H. & **De Boeck, P.** (2004). Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research*, 13, 363–394.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons, editor, *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197–223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263–287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6, 145–178.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4, 83–106.

Wolfe, E. W., Kao, C.-W. & **Ranney, M.** (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, 15, 465–492.